Visual Data Exploration and Analysis

Panelists and Contributors
Wes Bethel, LBNL (Co-Chair)
Randy Frank, LLNL
Sam Fulcomer, Brown University
Chuck Hansen, University of Utah (Co-Chair)
Ken Joy, UC Davis
Jim Kohl, ORNL
Don Middleton, NCAR

1. What Is Visualization?

Scientific visualization is the transformation of abstract information into images. It plays an integral role in the scientific process by facilitating insight through analysis into observed or simulated phenomena. Visualization as a discipline spans many research areas from computer science, cognitive psychology and even art. Yet the most successful visualization applications are created when close synergistic interactions between visualization experts and domain scientists are part of the algorithmic design and implementation process, leading to visual representations with clear scientific meaning. Visualization is used to explore, to debug, to present, to analyze, and to gain understanding. Visualization is literally everywhere. Images are present in this report, on television, on the Web, in books, journals, and magazines. The common theme is the ability to present information visually that is rapidly assimilated by human observers, and transformed into understanding or insight.

2. Impact of Visualization

As an indispensable part a modern science laboratory, visualization is akin to the biologist's microscope or the electrical engineer's oscilloscope. Whereas the microscope is limited to small specimens or use of optics to focus light, the power of scientific visualization is virtually limitless. Visualization provides the means to examine data that can be at galactic or atomic scales, or at any size in between. Moreover, unlike the traditional scientific tools for visual inspection, scientific visualization offers the means to create visual representations of abstract concepts that are otherwise unseeable. Trends in demographics or changes in levels of atmospheric CO_2 as a function of greenhouse gas emissions are familiar examples of such unseeable phenomena.

Over time, visualization techniques evolve in response to scientific need. Each scientific discipline has its "own language," verbal and visual, used for communication. The visual language for depicting electrical circuits is much different from the visual language for depicting theoretical molecules or trends in the stock market. No single visualization tool can serve for all science disciplines. Instead, visualization researchers work hand in hand with domain scientists as part of the scientific research process to define, create, adapt, and refine software that to incorporate domain knowledge and therefore "speak the visual language" of each scientific domain.

3. Research Areas

In this section we present a number of visualization research topics. They are a blend of computer science technologies for realizing needed growth in visualization capacity and capability, as well as new visualization technologies that address specific science needs. The challenges posed by modern computational science performed on large-scale computer systems are acute: not only is the amount of data becoming larger, but the complexity of the data itself is growing. Because of their fundamental design, visualization tools from earlier periods simply do not exhibit the capacity to process large scientific data sets. Similarly, the capabilities of earlier tools are not adequate to effectively present the meaningful information inherent in large, multidimensional data.

The topics discussed in this section take aim at known challenges posed by modern computational scientific research. Among these challenges are the fact that data sizes grow ever larger as computing capacity increases. With larger and more detailed simulation comes the need for more sophisticated visual analysis techniques, as well as the need for visualization infrastructure that provides the ability to simultaneously perform scalable data analysis that spans multiple data sets. Complicating matters even more is the case when multiple data sets are distributed across multiple sites. Even dramatic improvements in network technology cannot accommodate the "MxN" data movement required to aggregate data in this situation.

The most promising avenue for taking on large and distributed data visualization problems is parallel processing; task parallelism allocates specific tasks to processors in assembly-line fashion, and data parallelism spreads the workload for a single dataset across multiple processors. Both forms of parallelism require careful attention to design and implementation. Another challenge is the fact that as computing technology at large centers becomes more accessible to the research community, the remote user population will grow in size, and will expect more support for scalable tools that provide the ability to perform scientific research from remote locations: data must be analyzed where it was created without incurring the cost of large-scale data movement across the wide area network. Design and implementation of "traditional visualization software," as well as most commercial visualization products, have not taken into account these challenges, which in fact are tomorrow's requirements.

3.1 High-Capacity Visualization

One of the most significant challenges facing visualization is the need to process and display very large scientific datasets. Significant early advances by the visualization community in this area have identified areas requiring research to meet the needs of emerging computational science programs. One such area is data models and algorithms for processing and visualizing time-varying data, which add complexity to the large-data visualization challenge. Technologies that are already used to accelerate static data processing (such as multiresolution representations) can be applied with some degree of success to the access and processing of independent time steps of dynamic, time-varying data. However, new algorithms that effectively accelerate visualization of time-varying data, particularly out-of-core methods, are sorely needed.. New, related visualization technology that focuses on effective visual display of time-varying data will enable better

scientific understanding of complex dynamic phenomena. Achieving these objectives requires careful attention to the architecture of pipelined and parallel visualization processing tools, along with effective use of high-resolution displays.

We can draw a parallel between gains reasonably anticipated through improved processing of time-varying data and the gains realized through the same multiresolution techniques used in simulation codes themselves. The figure below depicts a multiresolution technique known as Adaptive Mesh Refinement (AMR). In AMR, a computational grid is locally refined to higher resolution in "regions of interest." A reactive chemistry combustion simulation would refine the computational grids in regions where there is a substantial amount of chemical reactivity, such as along a flame front. The primary benefit of AMR is that it is possible to achieve very high spatial and temporal resolutions that are local in scale; the cost of local refinement does not propagate to the full computational grid. Another example where substantial efficiency gains are realized through AMR is astrophysics simulations. In these codes, the range of spatial resolution in the computational domain varies from the cosmological or interstellar level down to planetary scales, where most of the volume in between is empty. The exact amount of efficiency gain is difficult to generally quantify since the refinement is a function of the particular phenomenon being modeled as well as parameters that specify the maximum amount of permissible error.

AMR codes typically realize somewhere between one and two orders of magnitude in efficiency gain compared to non-adaptive approaches. The gain in efficiency spans the entire processing pipeline, starting with storage of data on disk, continuing through data movement, downstream CPU cycles for analysis, including visualization. Use of AMR approaches for efficient representation of time varying data is not a new idea, but is one whose potential has not been realized due to lack of development. AMR is just one possible approach.

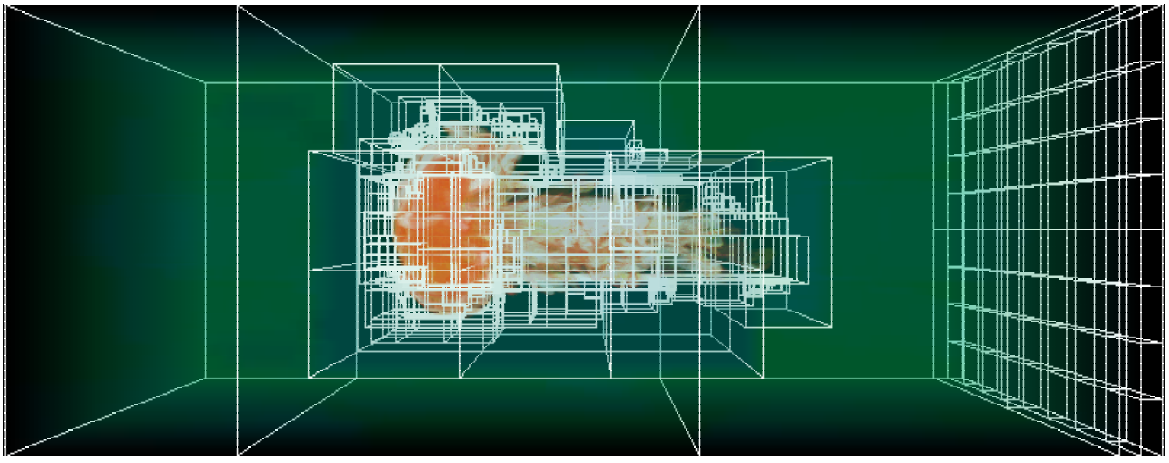

Figure 1. Direct Volume Rendering of Adaptive Mesh Refinement Data by Oliver Kreylos, LBNL and UC Davis. Argon Bubble Data courtesy of John Bell, LBNL.

3.2 Remote Visualization

Remote visualization is an integral part of all our lives. When we watch the weather forecast on television, we are viewing a presentation of data assembled from a number of remotely located sources: satellite imagery, regional ground-based stations, weather balloon observations, and computer simulations that predict tomorrow's weather. This same metaphor applies to modern computational science, where large datasets are generated on supercomputers and are analyzed or viewed by remotely located researchers. The trend toward consolidated centers that provide extreme computing capabilities as centralized resources, combined with the increased size of generated data, produce an acute need for remote visualization capabilities. As research teams are increasingly composed of geographically distributed scientists, interactive and collaborative remote visualization technologies can help to accelerate scientific discovery while reducing the costs associated with travel (see Fig. 2). There is an overlap between the needs of remote visualization and the objectives of other DOE research areas. A user should be required to authenticate only once in order to use a vast web of distributed resources, and they should expect that their data streams are adequately secure. As remote and distributed applications evolve, the ideal target is that suitable Grid infrastructure that supports single sign-on authentication and secure transmission of data streams is uniformly deployed across DOE facilities.

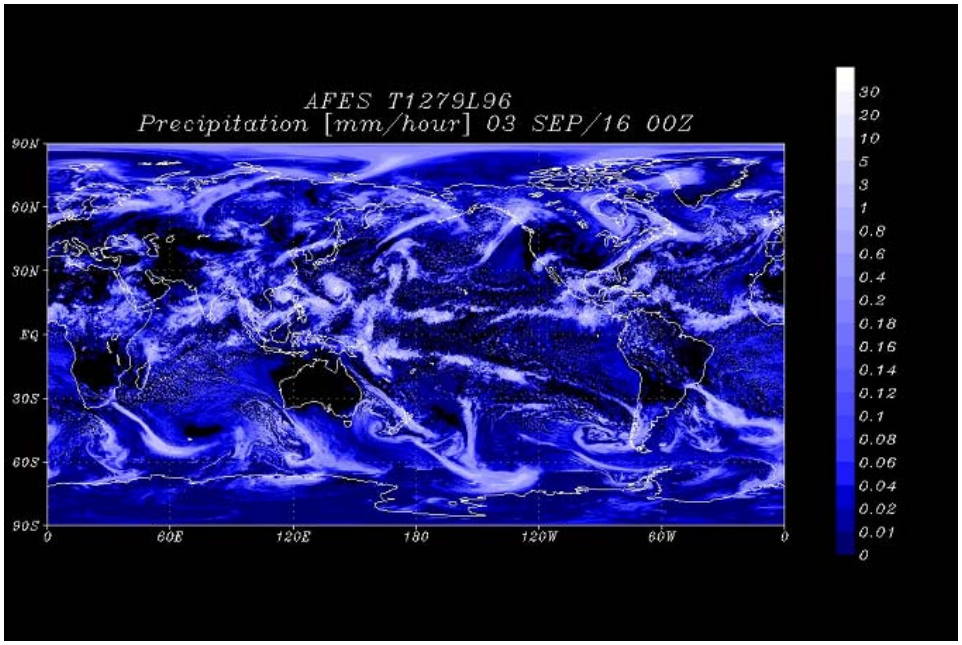

Figure 2. High-resolution datasets are computed at centralized facilities but are viewed by remotely located researchers. Remote visualization techniques help scientists make effective use of centralized facilities. *Don Middleton, NCAR.*

3.3 Multiresolution Methods

One avenue for addressing the problems posed by remote visualization is to enable the researcher to examine data at different resolutions. A quick examination of a low-resolution model or a statistical summary might reveal that no further inspection is necessary, thereby resulting in a significant time and resource savings. Alternatively, a

low-resolution model can provide a visual roadmap for high-resolution exploration, allowing a researcher to select small, high-resolution subsets of a dataset for more thorough analysis. Creating such multiresolution representations for specific scientific domains is a research area unto itself. However, creating effective methods for visually presenting such multiresolution representations and enabling the interactive transition between visual depictions are both active areas of visualization research (see Fig. 3).

Advances in data modeling technology will help to create statistically valid or bounded-error representations of fields that are more compact than the original. Such multiresolution techniques are important so that remote users may quickly examine simulation results, and have the option to "drill into" the raw, full-resolution data if desired. If possible, it is desirable to use techniques similar to, if not the same as, those used by the simulation itself. Adaptive Mesh Refinement is particularly attractive for it provides multiple levels of resolution that are scientifically significant.

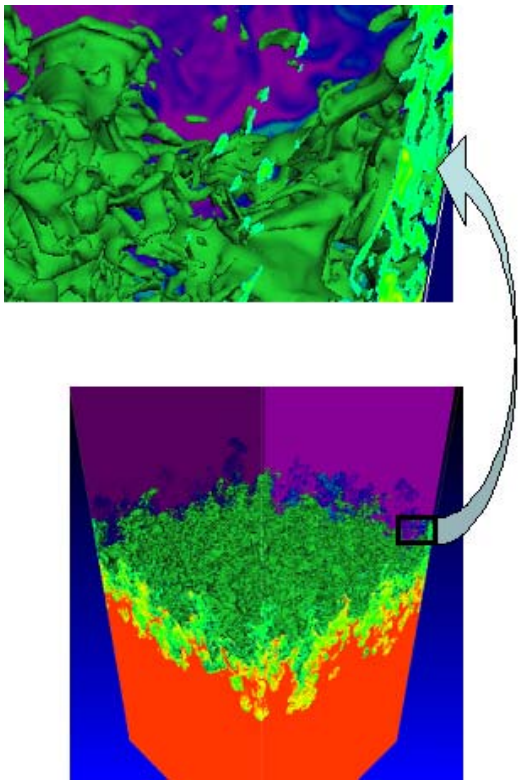

Figure 3. Multiresolution visualization requires specialized data models. *Randy Frank, LLNL.*

3.4 Multidimensional and Multivariate Visualization

Scientific computing has evolved to simulate phenomena at ever-increasing levels of fidelity and accuracy. Accurate modeling of phenomena often requires solving for more and more unknown variables. In order to facilitate scientific advances and provide insight into these complex systems, visualization technology is needed that can effectively display many variables simultaneously. The visualization challenge is compounded by

the scientific need for comparative analysis of experimental and simulation data, as well as data obtained or computed over a period of time. .

One approach to multivariate and multidimensional visual data analysis is based upon the idea of "data mining." In data mining, a user navigates between different datasets, or different resolutions of a dataset, based upon observations that in turn raise questions or spark ideas. Another scenario would leverage off-line analysis to locate and/or track domain specific features in the data to assist in data navigation. This area of research spans not only visual data analysis, but also includes the science of data summarization along with efficient storage and retrieval of large and diverse data.

3.5 Coupling Analysis, Visualization, and Data Management

At the core of the visualization processing pipeline is technology for accessing, manipulating, and processing data. As data models and data management systems evolve to accommodate ever-increasing dataset sizes and locations, there is a corresponding need for visualization tools to take advantage of these emerging technologies that store, retrieve, characterize, and analyze data. Statistical analysis forms an integral part of data understanding, yet few techniques exist for visualizing error, uncertainty, and other statistical features. Identification and characterization of interesting features are highly domain specific. Automatic detection and display of such features is a blend of statistical analysis, data management, and domain-specific visualization techniques. Through advances in visualization technology that include closer ties to data management technology (e.g., processing and display of statistical information), computational science programs benefit from increased visual data analysis capacity and capability.

3.6 "Behavioral" Visualization

As computer simulations increase in complexity, there is a growing need for visual representations of complex processes. One example is the behavior of optimization calculations in combinatorial algorithms. Visualization of algorithmic behavior, decision trees, and related "behavioral processes" provides insight into the operation and improvement of complex scientific software. A good example is how the search space in protein conformation is pruned to identify minimal energy conformations in complex molecules. Another example is the visual display of chemical pathways in combustion simulations, or metabolic pathways in cells (see Fig. 4). The evolution of simulation programs requires new visualization techniques to facilitate scientific insight .

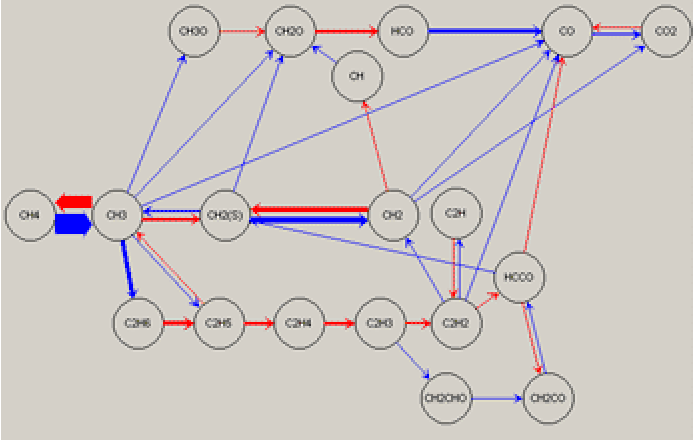

Figure 4. Chemical pathway visualization. The nodes represent species, and the edges represent flow of a conserved quantity, such as transfer of a particular element. *Mark Day, LBNL.*

4. Delivering Visualization Technology to Application Scientists

Application scientists have indicated that the best software tools are those specifically tailored for their domain. Such tools provide results in a familiar "language" that are readily comprehensible and applicable to scientific research (see Fig. 5). To develop such tools, visualization researchers must be part of the multidisciplinary science team performing the research. Even though each discipline needs tailored software tools, careful general-purpose software design and implementation will result in a "toolbox" of compatible components that can be combined in various ways to provide domain-specific solutions. Such components, with supporting data models, provide the "standards" to which disparate teams of visualization and science researchers can create compatible software tools. The evolution of a community-defined and supported software technology base will accelerate the growth of visualization research and its application to scientific domains through reduced duplication of effort and software engineering practices that promote reuse.

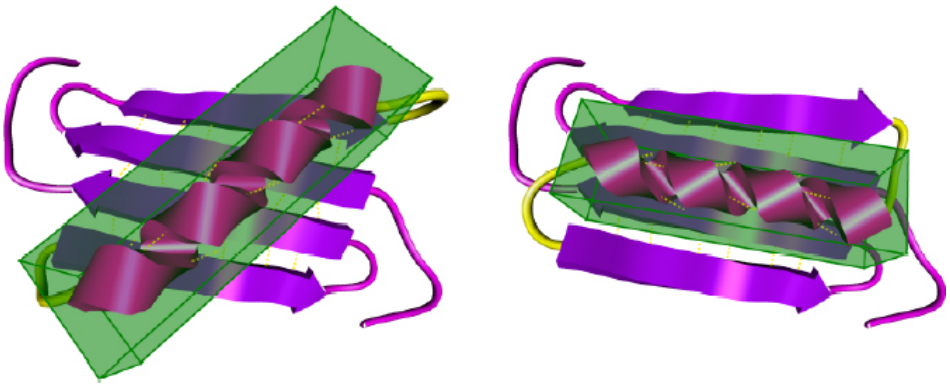

Figure 5. Visualization and manipulation of protein molecules is performed using "units" familiar to computational biologists – alpha helices and beta sheets. *Oliver Kreylos, UC Davis/LBNL; Silvia Crivelli, LBNL; Nelson Max, UC Davis, LLNL and LBNL. W. Bethel, LBNL, B. Hamann, UC Davis/LBNL.*

5. Resources Required (and Barriers Imposed)

The current model for funding visualization research and development tends to emphasize technology demonstrations. In contrast, science researchers need stable, production-quality software. The cost of ongoing software maintenance, documentation, training, and evolution far exceeds the cost of initial research and development. However, there is no funding mechanism to sustain these crucial activities. The traditional economic model of technology transfer from research into commercial products does not apply to scientific software, particularly visualization. The primary economic factor that results in a successful software commercialization – a large market that makes it possible to realize economies of scale – simply does not exist in the high performance computing world. Compared to traditional consumer markets for desktop publishing, photo editing, and so forth, the size of "the market" for high performance visualization software is very small. As such, software companies would be forced to charge a substantial fee for high performance visualization software. Not only are scientific research budgets tight, but when they are reduced, visualization is often the first line item that is cut. Additionally, vendors of commercial visualization software are faced with the inordinate task of porting and supporting their software products on an ever-changing array of computer hardware and software. Given the small market, the explicit conundrum between the commercial need to charge substantial fees and the research need to minimize expenditures, and the difficulty of maintaining a commercial product on a wide variety of platforms, commercial support for high performance visualization software is simply unattractive to industry. The most successful "commercial" visualization operations are those that produce an Open Source product, that invite community involvement in development, and which receive funding for ongoing development that targets the current needs of the research community. However, best-effort support often adds burden to visualization projects that depend on Open Source projects.

Scientific visualization also places extreme demands on computing infrastructure. All aspects of the computing pipeline are subject to significant demands for multiterabyte datasets: storage systems that serve as repositories; CPU and memory systems that process the data; networks that transport the data, and graphics systems that display it. The same maladies that plague the general scientific computing hardware market are present in the high-performance graphics and visualization world: the needs of the scientific visualization community are largely ignored by graphics hardware manufacturers. Those vendors are primarily drive by the needs of the computer gaming industry, which uses benchmarks that measure the number of frames per second generated when playing one of several different computer games. These ratings do not correlate to scientific visualization needs (see Fig. 6).

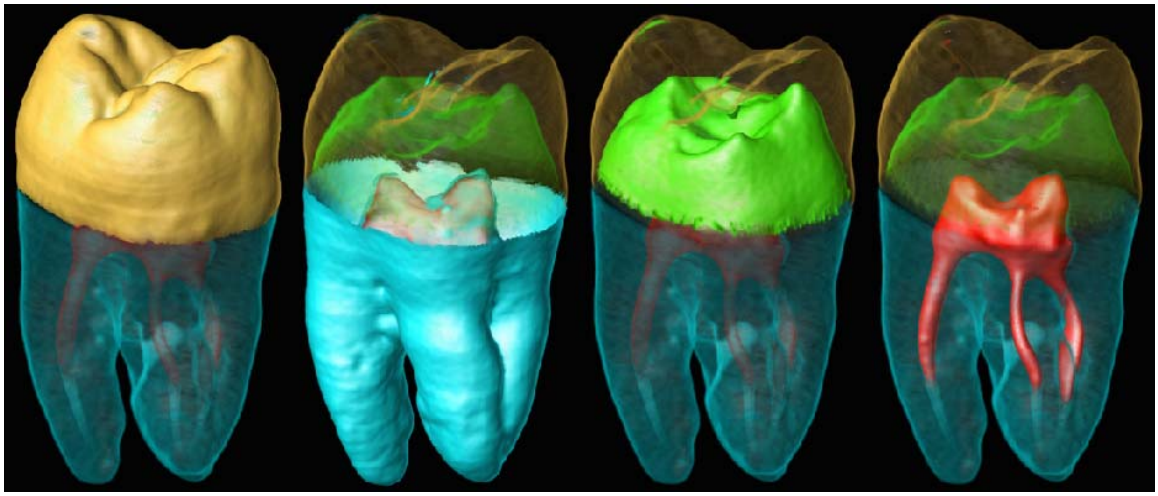

Figure 6. Advanced rendering features like three dimensional transfer functions are not provided by graphics hardware vendors because they aren't used by computer games. *Chuck Hansen, University of Utah.*

Given the central role of the remote visualization metaphor in modern scientific computing, there is an alarming lack of networking capacity to connect remote users with centralized facilities. Large-scale computer systems provide massive computational capacity but are often linked to the outside world using networks of inadequate capacity. Commodity Gigabit Ethernet hardware for desktop platforms is very inexpensive, yet the networks connecting major sites typically can support only two and a half such users operating at full capacity. Beyond the trunk lines themselves is the acute need for hardware that connects sites to the network. Effective use of centralized facilities requires high-speed network connectivity to deliver results to remotely located researchers. A difficult question is "how much networking capacity is required?" Like many of the questions raised throughout this document, the answer is multidimensional and highly dependent upon how the technology is to be used. In one view, the purpose of network backbones in a "Grid Computing" environment is to connect multiple, diverse resources so they all appear as one resource to the researcher. In this view, it is reasonable to say that the network should perform at a rate commensurate with the computational resources it connects. An approximate performance metric in this scenario calls for network performance that is in the range of tens of gigabits per second. Such networks are starting to come into existence now, as evidenced by the National Science Foundation's TeraGrid. Not only are fast networks needed, but the computational science requires that these many networks – commercial and those sponsored by advanced Federal research and development – are interconnected. Researchers need access to their data and computational resources, regardless of their location. A single network may provide adequate performance between a small number of sites, but researchers are realistically more dispersed, and may not be able to perform their work at one of the few sites endowed with adequate networking capacity. In other words, all federally funded networks should be "peered" so that a researcher at any federal research organization has outstanding network connectivity (OC-192 or 10Gb/s) to any other site. Funding streams from different organizations have inadvertently produced "islands" of network capacity.

When designing large-scale platforms, the needs of computational science research programs are taken into account by considering grid resolution, number of unknowns, number of time steps, and related variables to estimate the approximate amount of computing power required for a given class of algorithms. On the other hand, visualization processing is typically delegated to relatively small computing platforms that have nowhere near enough computing power. A disparity of several orders of magnitude in computing power is typical: simulations are run on platforms that can reach tens of teraflops, yet visualization is delegated to machines that are capable of only a few gigaflops. A substantial increase in funding for visualization computing platforms is critical to "impedance match" the capacity of simulation and analysis platforms. Similarly, an increase in visualization research staffing is needed to support projected growth trends to meet the needs of science research programs. In its early planning stages, the ASCI Program carefully defined visualization metrics that would be required to meet user needs given projected levels of computing capacity. Other sites and programs should adopt similar guidelines for future purchases. Otherwise, we can find ourselves in a situation similar to the Earth Simulator when the machine had to be idled so that storage and data analysis tasks were given an opportunity to "catch up."

6. Metrics of Success

Visualization success can be characterized by using several metrics. First and foremost is the degree to which visualization helps advance science as an enabling technology. The most obvious metric is the number of scientific discoveries facilitated by visualization. However, achieving these discoveries requires close coupling between visualization and scientific researchers so that visual data analysis tools are effectively designed and applied. Therefore, a practical programmatic objective would be to aim for an increase in the number of multidisciplinary teams where visualization is included. While such presence doesn't guarantee scientific discovery, it does create the potential for increased synergy as part of the scientific research process. Achieving such an increase of visualization in science can be implemented at the institutional level or at the individual project level. Another metric is longevity, or the temporal lifetime of visualization technology. The current visualization funding model encourages exploration of ideas but does not provide for the critical ongoing maintenance and lifecycle support activities needed to ensure that today's research prototypes form the basis for tomorrow's staple software tools. Increasing the lifetime of visualization technology will have long-term payoff in the form of reducing duplication of effort between visualization efforts. It will also simplify use of software tools since researchers will not be frequently required to surmount a steep learning curve associated with a new technology. Still another metric is the degree to which visualization, analysis, and data management are interoperable. Future research programs in visualization must include interoperability as a central theme to promote both longevity and widespread use by a large population.